Quantiles géométriques et sondage

Mohamed Chaouch & Camelia Goga

IMB, Université de Bourgogne, 9 Avenue Alain Savary, 21078 DIJON, France
mohamed.chaouch@u-bourgogne.fr, camelia.goga@u-bourgogne.fr

Résumé

Dans ce travail, nous nous sommes intéressés à l'estimation du quantile géométrique pour des données issues d'un plan de sondage. Nous donnons un estimateur du quantile géométrique basé sur le plan de sondage ainsi qu'une méthode itérative pour l'obtenir à partir des données d'échantillonnage. Sous des conditions générales, nous dérivons la variance asymptotique de l'estimateur du quantile et nous proposons un estimateur convergent de cette variance. Le bon comportement de l'estimateur du quantile géométrique est vérifié par une étude par simulation.

Mots-clés : Représentation de type Bahadur, estimateur convergent, équation estimante, estimateur de type Horvitz-Thompson, estimation de la variance.

Abstract

In this work, we are interested in estimating geometric quantile when data are obtained in a complex survey. We construct a design-based estimator of the geometric quantile and compute it by iterative method from survey data. Under broad assumptions, we derive the asymptotic variance of the quantile estimator and propose a consistent estimator of it. Finally, the good behavior of the geometric quantile estimator is verified through a simulation study.

Key words: Bahadur expansion, consistent estimator, estimating equation, Horvitz-Thompson estimator, variance estimation.

1 Introduction

Les quantiles univariés, conditionels ou non conditionnels, sont fréquemment utilisés en Statistique. Par exemple, la médiane est un indicateur robuste de la tendance centrale d'une population, l'intervalle interquartile est un bon indicateur de sa dispersion. En pratique, ces quantiles sont calculés suivant un critère d'ordre sur les observations. L'ordre n'étant pas total sur $\mathbb{R}^d$, une extension de la définition classique des quantiles au cas où les observations sont à valeurs dans $\mathbb{R}^d$, avec $d \geq 2$, ne peut être que partielle. Il s'agit dans

ce cas du vecteur quantile (dit *"arithmétique"*) dont les composantes sont les quantiles marginaux. Cette définition souffre de plusieurs faiblesses.
Dans les dernières années, plusieurs auteurs se sont intéressés à la généralisation de la notion de quantiles dans le cadre de variables aléatoires multidimensionnelles. Deux approches principales ont été développées : la première approche est basée sur la notion de fonction de profondeur (voir Liu et *al.*, 1999), la seconde approche définie les quantiles multivariés comme étant des M-estimateurs qui minimisent une fonction de perte ou de coût. Dans la suite nous nous focalisons sur la définition des quantiles, dit *géométriques*, introduite par Chaudhuri (1996) qui correspond à une des définitions qui a été proposée dans le cadre de la deuxième approche.
La manière dont les données sont obtenues est rarement prise en compte dans l'estimation des quantiles géométriques et on suppose souvent que les observations sont indépendantes et identiquement distribuées. Or cette hypothèse n'est pas systématiquement vérifiée. D'autre part, Chaudhuri (1996) a proposé un algorithme permettant de calculer l'estimateur du quantile géométrique dont le temps de calcul dépend de la taille de l'échantillon. Pour ces deux raisons, nous proposons d'utiliser des techniques de sondage pour estimer le quantile géométrique.
Nous commençons tout d'abord par adapter la définition des quantiles géométriques au cadre de sondage, ensuite nous définissons un estimateur. En utilisant la technique de linéarisation par les équations estimantes, nous donnons une représentation de type Bahadur de notre estimateur qui permettera par la suite de déduire une approximation de sa variance. Une analyse par simulation a été faite pour montrer le bon comportement de nos estimateurs.

2 Quantiles géométriques et sondage

On considère une population $U = \{1, 2, \ldots, N\}$ de taille finie N. On s'intéresse à une variable déterministe $Y \in \mathbb{R}^d$ définie pour chaque individu k de la population U. Considérons la fonction de perte *multivariée* définie par

$$\phi(\mathbf{u}, \mathbf{t}) = ||\mathbf{t}|| + \langle \mathbf{u}, \mathbf{t} \rangle ,$$

avec $\mathbf{t} \in \mathbb{R}^d$ et $\mathbf{u} \in B^d = \left\{ \mathbf{u} \in \mathbb{R}^d \ : ||\mathbf{u}|| < 1 \right\}$.

Le quantile géométrique d'ordre u, du nuage des points $Y_1, \ldots, Y_N$ dans $\mathbb{R}^d$, minimise le total de la fonction de perte ϕ calculé sur toute la population U,

$$Q_N(u) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{k=1}^{N} \phi(u, Y_k - \theta) . \tag{1}$$

En utilisant les mêmes arguments que Kemperman (1987), nous pouvons montrer que $Q_N(u)$ existe et est unique lorsque l'hypothèse suivante est vérifiée :

($\star$) $\{Y_1, \ldots, Y_N\}$ ne sont pas alignés dans $\mathbb{R}^d$.

Nous pouvons déduire à partir de la relation (1) que $Q_N(u)$ est l'unique solution de l'équation suivante dont l'inconnue est θ :

$$\sum_{k=1}^{N} \left[S\left(Y_k - \theta\right) + u \right] = 0 \tag{2}$$

où $S(v) = v/\|v\|$, pour tout $v \in \mathbb{R}^d$ tel que $v \neq 0$. Notons par $h_k(\theta) = S(Y_k - \theta) + u$, pour tout $k \in U$, et par $H_U(\theta) = \sum_{i=1}^{N} h_k(\theta)$ le total de $h_k(\theta)$ calculé sur toute la population U. Notons que le quantile $Q_N\left(u\right)$ vérifie l'équation $H_U(Q_N(u)) = 0$.

2.1 Echantillonnage et estimation

Un échantillon s, *i.e.* une partie $s \subset U$, est tiré selon un procédé probabiliste $p(s)$ où p est une loi de probabilité sur l'ensemble de parties possibles de U. On note $\pi_k = Pr(k \in s)$ pour tous les $k \in U$ et $\pi_{kl} = Pr(k \ \& \ l \in s)$ pour tous $k, l \in U$, $k \neq l$ les probabilités d'inclusion du premier et deuxième degré. On suppose par ailleurs que $\pi_k > 0$ et $\pi_{kl} > 0$: tous les individus et les couples d'individus ont une probabilité non-nulle d'être présents dans l'échantillon.

Un estimateur par substitution du quantile géométrique $Q_N(u)$ peut s'écrire de la façon suivante

$$\widehat{Q}\left(u\right) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{k \in s} \frac{\phi\left(u, Y_k - \theta\right)}{\pi_k}. \tag{3}$$

Nous pouvons démontrer que $\widehat{Q}\left(u\right)$ existe et unique lorsque l'hypothèse suivante est vérifiée :

($\star\star$) Supposons que les $Y_k \in \mathbb{R}^d$, pour tout $k \in s$, ne sont pas alignés dans $\mathbb{R}^d$.

Dans la suite nous donnons les propriétés asymptotiques de cet estimateur. Pour cela notons d'abord par $\widehat{H}(\theta) = \sum_{k \in s} \frac{h_k(\theta)}{\pi_k} = \sum_{k \in U} \frac{h_k(\theta)}{\pi_k} I_k$, l'estimateur de type Horvitz-Thompson (1952) de $H_U(\theta)$.

Nous savons que $\mathbb{E}_p(I_k) = \pi_k$, par conséquent, $\mathbb{E}_p(\widehat{H}(\theta)) = H_U(\theta)$, où $\mathbb{E}_p(\cdot)$ désigne l'espérance par rapport au plan de sondage. Nous pouvons également écrire la variance de type Horvitz-Thompson de $\widehat{H}(\theta)$ comme suit

$$\mathbb{V}_p\left(\widehat{H}(\theta)\right) = \sum_{U} \sum_{U} \Delta_{k\ell} \frac{h_k(\theta)}{\pi_k} \frac{h_l^T(\theta)}{\pi_\ell}. \tag{4}$$

Un estimateur de type Horvitz-Thompson non biaisé de $\mathbb{V}_p\left(\widehat{H}(\theta)\right)$ est donné par l'expression suivante : $\widehat{\mathbb{V}}_p\left(\widehat{H}(\theta)\right) = \sum_s \sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{h_k(\theta)}{\pi_k} \frac{h_l^T(\theta)}{\pi_\ell}$.
Nous pouvons démontrer le résultat suivant :

Théorème 1. *Soit $u \in B^d$ et $\widehat{Q}(u)$ l'estimateur de $Q_N(u)$ calculé à partir de l'échantillon s.*

1. *Si $\widehat{Q}(u) = Y_k$ pour un certain $k \in s$, alors* $\| \sum_{\substack{k\in s \\ Y_k \neq \widehat{Q}(u)}} \frac{h_k(\widehat{Q}(u))}{\pi_k} \| \leq (1+\|u\|) \sum_{\substack{k\in s \\ Y_k = \widehat{Q}(u)}} \frac{1}{\pi_k}$

2. *Si $\widehat{Q}(u) \neq Y_k$ pour tout $k \in s$, alors*

$$\widehat{H}(\widehat{Q}(u)) = \sum_{k\in s} \frac{h_k(\widehat{Q}(u))}{\pi_k} = 0. \tag{5}$$

Ce théorème nous a permis de définir un algorithme de calcul de l'estimateur du quantile géométrique. Cet algorithme est consitué de deux étapes : la première étape consiste à vérifier, pour chaque individu appartenant à l'échantillon, l'inégalité donnée par le résultat (1) du théorème. Si un point Y_k vérifie l'inégalité, alors $\widehat{Q}(u) = Y_k$ sinon on passe à la deuxième étape de l'algorithme qui consiste à résoudre l'équation (5) en utilisant par exemple l'algorithme de Newton-Raphson.

2.2 Cadre asymptotique

L'estimateur $\widehat{Q}(u)$ est non linéaire, pour cette raison nous utilisons la technique de linéarisation pour obtenir une approximation de sa variance. Plaçons nous maintenant dans le cadre asymptotique du sondage (Isaki & Fuller, 1982) pour démontrer les différents résultats asymptotiques. On suppose que les hypothèses suivantes sont vérifiées :

(A1) $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0,1)$,

(A2) $\min_k \pi_k \geq \lambda_1$, $\min_{k\neq l} \pi_{kl} \geq \lambda_2$ avec λ_1, λ_2 deux constantes positives et $\overline{\lim}_{N\to\infty} n \max_{k\neq l} |\pi_{kl} - \pi_k\pi_l| < \infty$.

(A3) On suppose qu'il existe une constante positive M tel que $||Y_k - \theta|| \geq M$ pour tout $k \in U$ et $\theta \in \mathcal{V}_{Q_N(u)}$, où $\mathcal{V}_{Q_N(u)}$ désigne un voisinage de $Q_N(u)$.

(A4) $\widehat{Q}(u)$ est un estimateur convergent de $Q_N(u)$, c-à-d pour tout $\varepsilon > 0$ fixé, nous avons $\lim_{N\to\infty} \mathbb{P}\left(||\widehat{Q}(u) - Q_N(u)|| > \varepsilon\right) = 0$.

(A5) $\dfrac{\sqrt{n}}{N}\left[\widehat{H}(Q_N(u)) - H_N(Q_N(u))\right] \longrightarrow N(0,\Sigma)$ avec Σ une matrice définie positive.

Lemme 1. *Sous les hypothèses (A1) et (A2), l'estimateur de type Horvitz-Thompson, $\widehat{H}(\theta)$, de $H_U(\theta)$, satisfait $\mathbb{E}_p\left|\left|\frac{1}{N}\left(\widehat{H}(\theta) - H_N(\theta)\right)\right|\right| = O(n^{-1/2})$ pour tout $\theta \in \mathcal{V}_{Q_N(u)}$.*

Notons par $J_U(\theta)$ la matrice Jacobienne de $H_U(\theta)$ définie par

$$J_U(\theta) = \sum_U \frac{1}{||Y_k - \theta||}\left[I_d - S(Y_k-\theta)S^T(Y_k-\theta)\right],$$

avec I_d la matrice identité de dimension d. L'estimateur de type Horvitz-Thompson de $J_U(\theta)$ est

$$\widehat{J}(\theta) = \sum_s \frac{1}{\pi_k||Y_k - \theta||}\left[I_d - S(Y_k-\theta)S^T(Y_k-\theta)\right].$$

Les deux matrices $J_U(\theta)$ et $\widehat{J}(\theta)$ sont de dimension $d\times d$, symétriques et définies positives sous les hypothèses $(\star)$ et $(\star\star)$.

Lemme 2. Supposons que les conditions (A1)-(A3) sont vérifiées. *Pour tout $\theta \in \mathcal{V}_{Q_N(u)}$, nous avons*

(i) $\frac{1}{N}J_U(\theta) = O(1)$,

(ii) $\mathbb{E}_p\left|\left|\frac{1}{N}\left(\widehat{J}(\theta) - J_U(\theta)\right)\right|\right|_1 = O(n^{-1/2})$ *où* $||\cdot||_1$ *est la norme trace telle que* $||A||_1^2 = tr(A^TA)$ *pour toute matrice* A.

Théorème 2. *Lorsque les hypothèses (A1)-(A5) sont vérifiées, l'estimateur $\widehat{Q}(u)$ de $Q_N(u)$ basé sur le plan de sondage $p(s)$ satisfait la relation suivante :*

$$\begin{aligned}\widehat{Q}(u) - Q_N(u) &= -J_U^{-1}(Q_N(u))\left(\widehat{H}(Q_N(u)) - H_U(Q_N(u))\right) + o_p(n^{-1/2}) \\ &= \sum_s \frac{u_k}{\pi_k} + o_p(n^{-1/2})\end{aligned}$$

où $u_k = -J_U^{-1}(Q_N(u))h_k(Q_N(u))$ est la variable linéarisée de $Q_N(u)$ avec $\sum_U u_k = 0$. Par conséquent la variance asymptotique de $\widehat{Q}(u)$ notée $\mathbb{AV}_p(\widehat{Q}(u))$, est égale à la variance de l'estimateur $\sum_s \frac{u_k}{\pi_k}$,

$$\mathbb{AV}_p(\widehat{Q}(u)) = \sum_U\sum_U \Delta_{kl}\frac{u_k}{\pi_k}\frac{u_l^T}{\pi_l}.$$

La variance asymptotique de $\widehat{Q}(u)$ est calculée sur la population entière U, alors qu'on ne dispose que d'un échantillon s de cette population. Nous proposons maintenant de l'estimer par l'estimateur de type Horvitz-Thompson de la variance en remplaçant u_k par son estimateur $\widehat{u}_k$. Nous obtenons alors

$$\widehat{\mathbb{V}}_p(\widehat{Q}(u)) \;=\; \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\widehat{u}_k}{\pi_k} \frac{\widehat{u}_l^T}{\pi_l} = \left[\widehat{J}(\widehat{Q}(u))\right]^{-1} \widehat{V}_p(\widehat{H}(\widehat{Q}(u))) \left[\widehat{J}(\widehat{Q}(u))^T\right]^{-1}$$

avec $\widehat{u}_k = -\widehat{J}^{-1}(\widehat{Q}(u))h_k(\widehat{Q}(u))$. Le résultat suivant donne, sous certaines hypothèses, la convergence de $\widehat{\mathbb{V}}_p(\widehat{Q}(u))$ vers $\mathbb{AV}_p(\widehat{Q}(u))$.

Théorème 3. *Supposons que* $(A1)$-$(A5)$ *sont vérifiées et que de plus* $\frac{1}{N^2}\left[\widehat{V}_p(\widehat{H}(Q_N)) - \Sigma\right] = o_p(n^{-1})$, *alors* $\widehat{V}_p(\widehat{Q}(u)) - \mathbb{AV}_p(\widehat{Q}(u)) = o_p(n^{-1})$.

Implémentation informatique et simulations

Soit Y un vecteur bidimentionnel qui suit la distribution binormale $\mathcal{N}_2((0,0); I_2)$. Nous avons simulé une population de taille 5000 selon cette loi. Nous avons vérifié l'efficacité de l'estimateur du quantile géométrique d'un vecteur Y dans une étude par simulation et pour deux plans de sondage, le Sondage Aléatoire Simple sans remise (SAS) et le sondage stratifié. Pour deux directions u fixées nous avons évalué la qualité de nos estimateurs en calculant l'erreur relative moyenne relative à chaque plan de sondage. L'estimateur proposé fonctionne numériquement bien et sans surprise Il s'avère que le plan stratifié permet d'obtenir de meilleures estimations que le plan SAS.

Bibliographie

[1] Chaudhuri, P. (1996) On a geometric notation of quantiles for multivariate data. *Journal of the American Statistical Association*, 91, 862–872.

[2] Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.

[3] Isaki, C. I. and Fuller, W. A. (1982) Survey Design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89–96.

[4] Kemperman, J. H. B. (1987) The median of a finite measure on a Banach space. *Statistical Data Analysis based on the L_1-norm and related methods*, Y. Dodge (ed), North-Holland, Amsterdam, 217-230.

[5] Liu, R. Y., Parelius, J. M. and Singh, K. (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *The Annals of Statistics*, 27, 783–858.